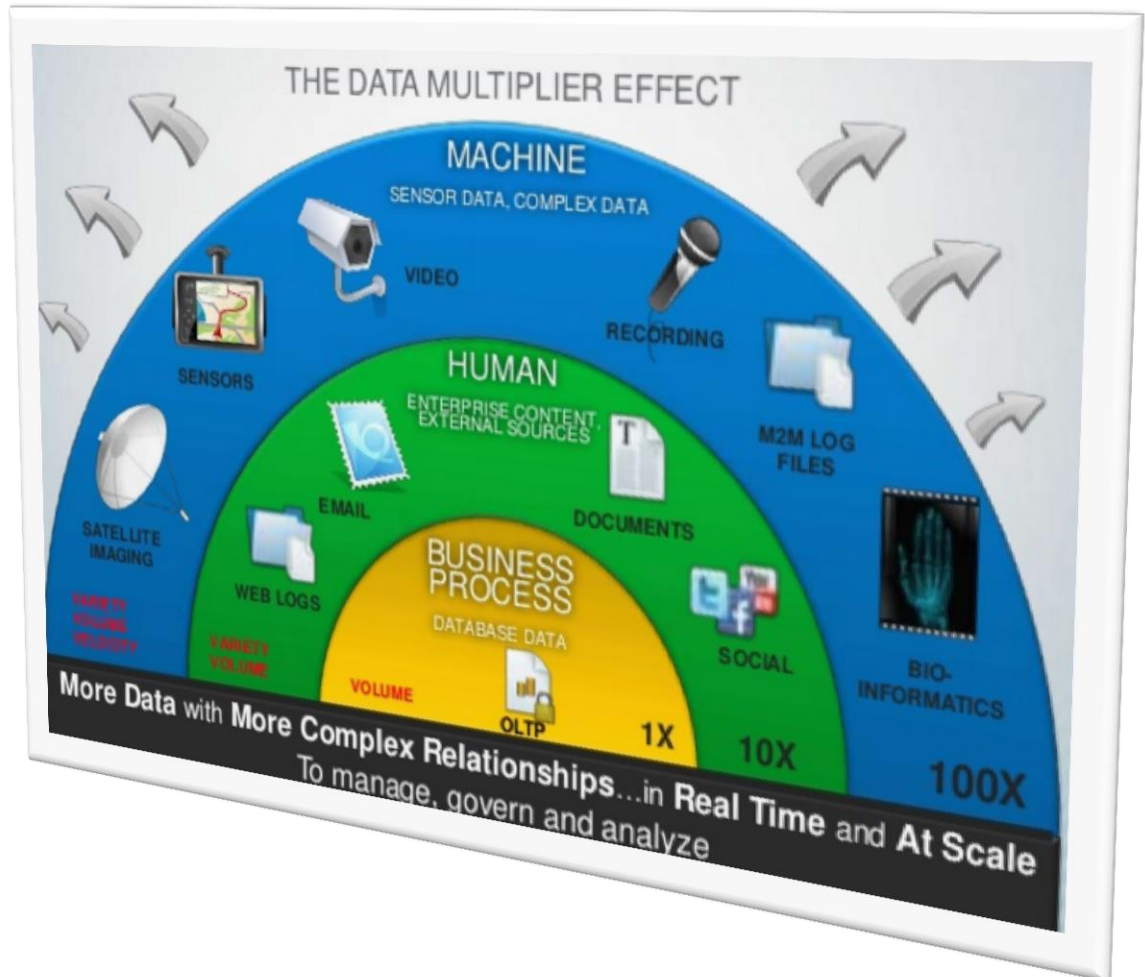
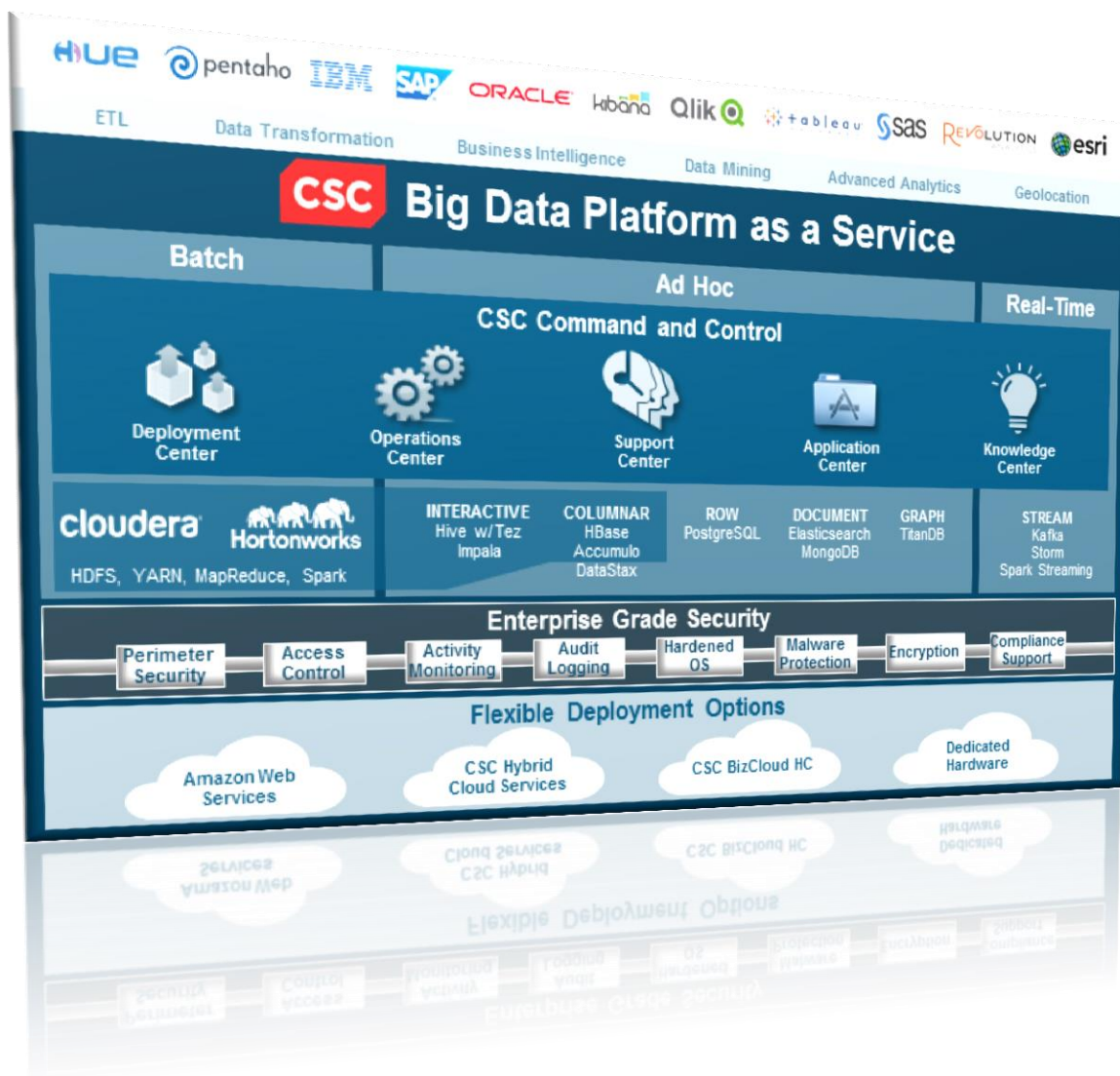


Big Data Analytics & IBM BIG INSIGHT

En la actualidad se generan grandes volúmenes de datos de diversos tipos, a gran velocidad y con diferentes frecuencias.



Las tecnologías disponibles permiten efectuar su almacenamiento, adquisición, procesamiento y análisis utilizando métodos y técnicas diversas. Es importante tener en cuenta que cuando se procesan y almacenan grandes volúmenes de datos entran en juego dimensiones adicionales a las técnicas, como el gobierno, la seguridad y las políticas. Por lo que desarrollar una solución de Big Data es un proceso complejo que requiere considerar muchos factores de acuerdo con cada problema particular. Durante los últimos años, la manera en que las personas acceden los datos y sus orígenes han cambiado drásticamente, como resultado del avance tecnológico.



Ahora bien, esta revolucion tecnologica genera grandes cantidades de datos que provienen de diferentes origenes, llamese estos redes sociales, sistemas tecnologicos o paginas web utilizadas, para marketing como Amazon. Por esta razon, Big Data es un concepto que hace referencia a grandes cantidades de informacion disponibles en diferentes formatos y tipos de estructuras recopiladas principalmente a traves de Internet mediante la interaccion de usuarios de computadores, telefonos moviles, dispositivos GPS entre otros.

Arquitectura de Big Data

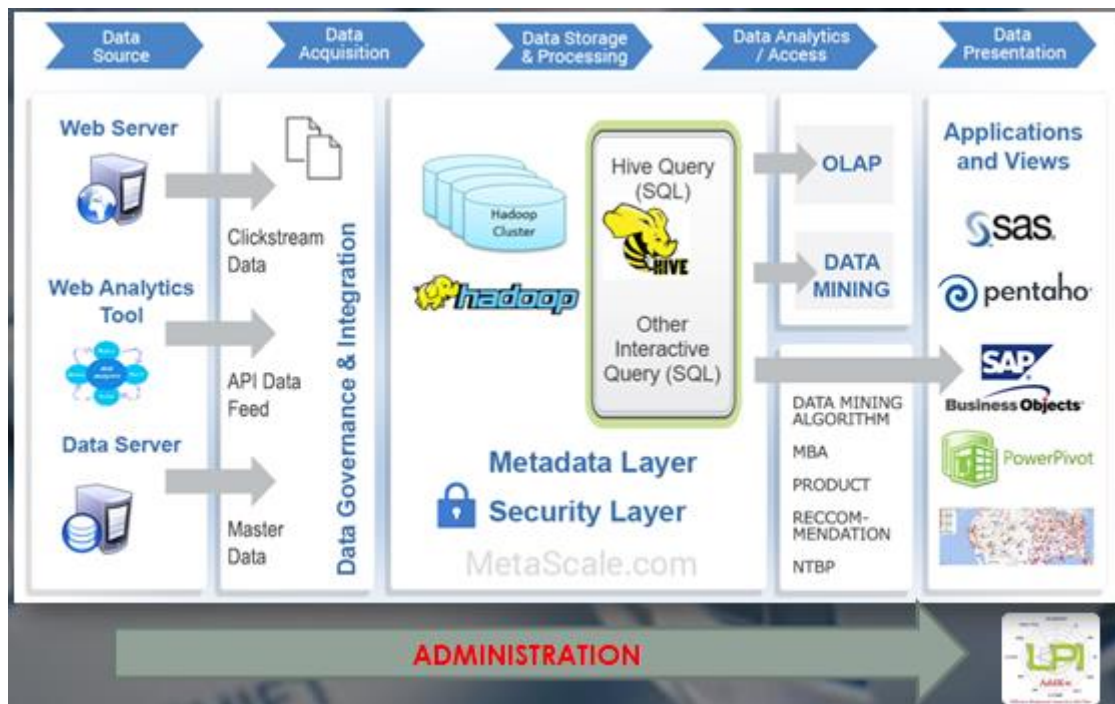
En terminos generales, una arquitectura de Big Data (Chunmei Duan¹, 2014) esta compuesta por cinco componentes: recoleccion de datos, almacenamiento, procesamiento de datos, visualizacion y administracion. Ademas, cada uno de estos componentes ha ido anadiendo nuevas tecnologias, las cuales dependen de las necesidades que se den y tambien su adaptacion es obligatorio para dar una solucion eficiente a las empresas actualmente.

Recoleccion de datos: Esta etapa se refiere a la obtencion de los datos. El sistema se conecta a las diferentes fuentes de informacion para obtener los datos que luego seran procesados, almacenados y posteriormente analizados. A continuacion, se describen los dos metodos de recoleccion de datos que se pueden dar y que dependen del caso de uso por desarrollar segun el criterio:

- Batch o por lotes: Este tipo de recoleccion se conecta cada cierto tiempo a las fuentes de informacion, llamese estas: sistemas de ficheros o Base de Datos; en las cuales se buscan cambios realizados desde la ultima conexion que se hizo.

– Streaming (clickstream) o por transmisión en tiempo real: Este tipo de recolección trabaja directamente con la fuente de información de manera continua de forma que la información se obtiene cada vez que se tramita (tiempo real).

Gracias a la evolución tecnológica, los sistemas de la actualidad pueden trabajar obteniendo la información de las dos formas usando streaming o batch. Asimismo, los sistemas modernos permiten filtrar la información; por ejemplo, según la información o formato que se quiere recolectar.



Almacenamiento

Actualmente, los sistemas de almacenamiento han tenido que adaptarse o buscar nuevas formas de almacenar su información debido a las grandes cantidades de información que se generan además de la velocidad con que se mueven. Por esta razón, los métodos de almacenamiento tradicionales como las Bases de Datos relacionales se han quedado cortos a la hora de tratar esta información conocida como Big Data.

Procesamiento de datos

Este paso presenta uno de los puntos más importantes a la hora de hablar sobre Big Data ya que una vez que se tienen almacenados los datos, se busca obtener conocimiento o valor por medio del procesamiento y análisis de toda esta información almacenada. Actualmente, se cuenta con herramientas muy poderosas que permiten procesar esta información que muchas veces presenta diferentes tipos de formato y orígenes como las bases de datos NoSQL o los sistemas de archivos. "(Chunmei Duan1, 2014)".

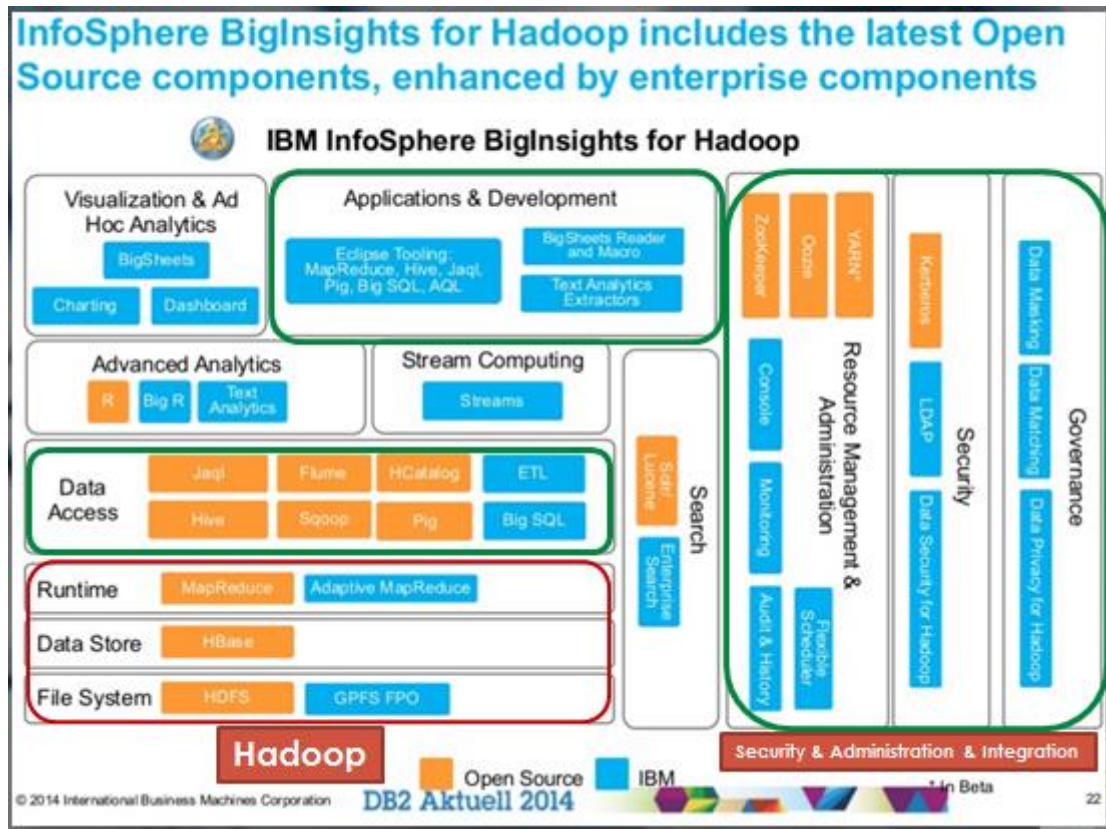
Visualización y Administración

Esta capa de Big Data, muestra el producto del almacenamiento y procesamiento de la información que da como resultado la producción de conocimiento.

Arquitectura de IBM: BigInsights

BigInsights es una plataforma de software para descubrir, analizar y visualizar datos de diferentes orígenes. Este software es normalmente utilizado para ayudar a procesar y analizar el volumen, variedad y velocidad de datos que entra continuamente a las diferentes organizaciones cada día. Asimismo, es una plataforma flexible construida sobre el Framework de código abierto Apache Hadoop que es ejecutado en paralelo sobre hardware de bajo costo. Esta solución de IBM ayuda a los desarrolladores, científicos de datos y administradores en las organizaciones a rápidamente construir y desplegar análisis personalizados de información mediante el procesamiento de los datos.

Estos datos son a menudo integrados en las bases de datos existentes, almacenes de datos (Data Warehouses) y la infraestructura de inteligencia de Negocios. Además, mediante el uso de BigInsights, los usuarios pueden extraer nuevos conocimientos a partir de estos datos para mejorar el conocimiento de su negocio.



Características y arquitectura de BigInsights

Esta plataforma brinda un conjunto de herramientas o componentes tecnológicos que brindan las capacidades necesarias para que una organización pueda procesar los grandes volúmenes de datos recibidos diariamente. En la figura a continuación, se muestran los diferentes componentes que forman la arquitectura de BigInsights:

BigInsights extiende el Framework de Hadoop con seguridad a nivel empresarial, administración, disponibilidad, integración con almacenes de datos (Data Warehouses) existentes, además de herramientas que simplifican la productividad de los desarrolladores.

Hadoop ayuda a las empresas a aprovechar los datos que antes era difícil de manejar y analizar. BigInsights cuenta con Hadoop y sus tecnologías relacionadas como elemento fundamental. A continuación describimos algunos de los componentes utilizados en la arquitectura:

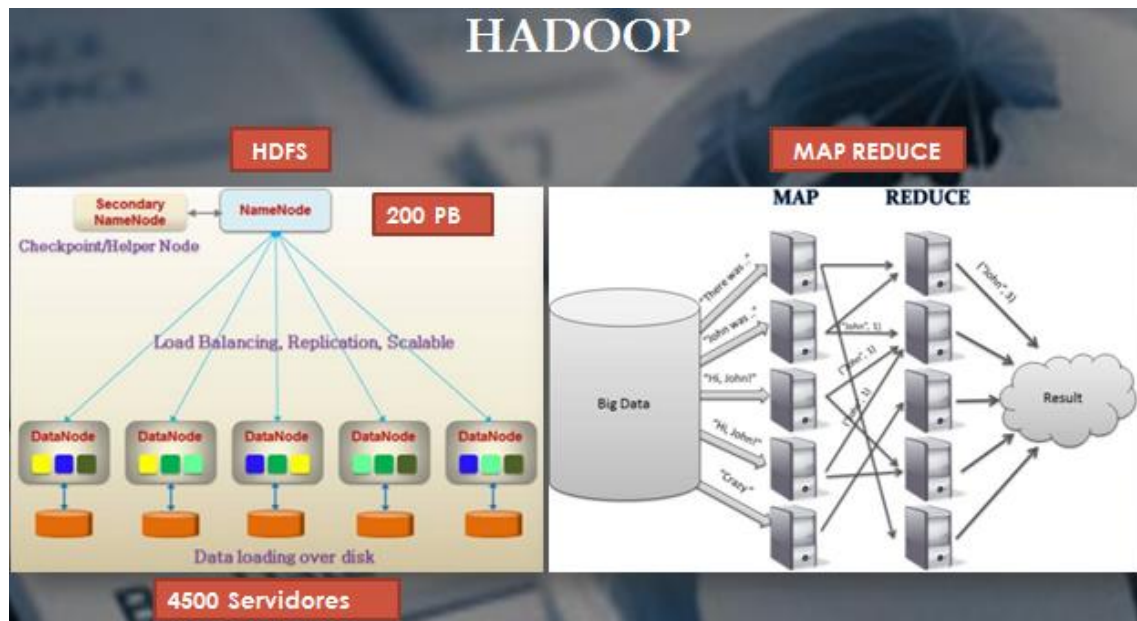
Hadoop en forma general procesa ingentes cantidades de datos a través de servidores conectados a nodos centrales (HDFS) y procesamiento en paralelo de estandarización de datos (MAPReduce).

HDFS: HDFS viene con la plataforma abierta de IBM (IBM Open Platform) mediante el uso de Apache Hadoop como su sistema de archivos distribuido.

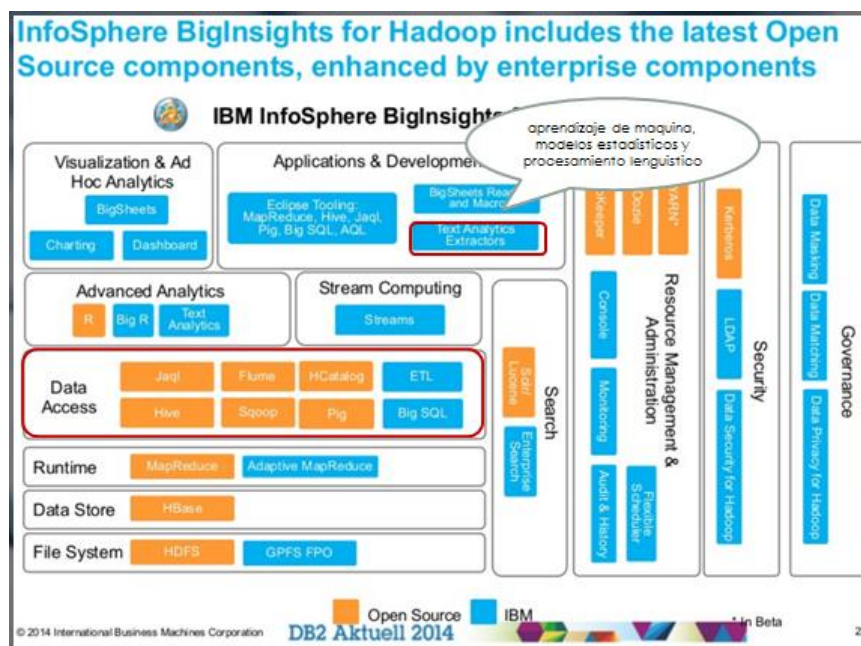
HDFS es un sistema de archivos desarrollado en JAVA que administra grandes cantidades de datos en forma escalable y confiable y que puede abarcar grandes grupos de servidores básicos. HDFS ha

demostrado un alto desempeño y escalabilidad de hasta 200 PB de almacenamiento en un cluster simple de 4500 servidores que soporta un billon de archivos y bloques.

MapReduce: El Framework MapReduce es el nucleo de Apache Hadoop. Este paradigma de programacion prevee escalabilidad masiva a traves de cientos o miles de servidores en un cluster Hadoop. MapReduce se emplea en la resolución práctica de algunos algoritmos susceptibles de ser paralelizados. Por regla general se abordan problemas con datasets de gran tamaño, alcanzando los petabytes de tamaño. Es por esta razón por la que este framework suele ejecutarse en sistema de archivos distribuidos (HDFS).



Tecnologías de código abierto: Hadoop incluye muchas tecnologías de código abierto y sus dependencias que continúan aumentando a medida que se utiliza Hadoop en más aplicaciones. Se emplean estas tecnologías para interactuar con el ecosistema de Hadoop.



Text Analytics: BigInsights incluye Text Analytics, que extrae información estructurada a partir de datos no estructurados y semi-estructurados.

Minería de textos incluye una serie de técnicas de aprendizaje de máquina, modelos estadísticos y procesamiento lingüístico que estructura la información que proviene de archivos de textos.

IBM BigSQL: BigSQL aprovecha la fuerza de IBM en los motores de SQL para proporcionar acceso ANSI SQL a los datos a través de cualquier sistema de Hadoop, vía JDBC o ODBC y obteniendo los datos de Hadoop o una base de datos relacional.

Integración con otros productos de IBM

BigInsights complementa y amplía las capacidades de negocio existentes mediante la integración con otros productos de IBM. Estos puntos de integración extienden las tecnologías existentes para abarcar más tipos de información, lo que permite una visión completa de su negocio. Algunos productos de IBM con los cuales se puede trabajar a la hora de usar Hadoop son los siguientes:

- IBM DB2.
- IBM Cognos Business Intelligence.
- InfoSphere Streams.
- IBM Netezza.
- IBM InfoSphere Data Explorer.

